

AD-A083 521

CLEMSON UNIV S C DEPT OF MATHEMATICAL SCIENCES  
RIDGE ESTIMATION IN LINEAR REGRESSION.(U)  
OCT 76 J S HAWKES

F/6 12/1

N00014-75-C-0451

UNCLASSIFIED

N80

NL

1 of 1  
AD  
A083521



END  
DATE  
FILMED  
5-80  
DTIC

ADA083521

DISTRIBUTION STATEMENT A

Approved for public release;  
Distribution Unlimited

(12) LEVEL II

6 RIDGE ESTIMATION IN  
LINEAR REGRESSION.

(10) JAMES S. / HAWKES /

DTIC  
ELECTE  
S APR 28 1980 D

14  
(11) REPORT N80 / 71-132 /  
OCT 1976 /  
(9) TECHNICAL REPORT #232  
(12) 11

DISTRIBUTION STATEMENT A

Approved for public release;  
Distribution Unlimited

THE AUTHOR'S WORK WAS SUPPORTED BY THE OFFICE OF NAVAL RESEARCH  
UNDER CONTRACT / N00014-75-C-0451 / ✓

4 182

## RIDGE ESTIMATION IN LINEAR REGRESSION

James S. Hawkes  
Clemson University

### ABSTRACT

Consider the linear regression model  $Y = X\theta + \epsilon$ . Recently, a class of estimators, variously known as ridge estimators, has been proposed as an alternative to the least squares estimators in the case of collinearity, that is, when the design matrix  $X'X$  is nearly singular. The ridge estimator is given by  $\hat{\theta} = (X'X + KI)^{-1} X'Y$ , where  $K$  is a constant to be determined. An optimal choice of the value of  $K$  is not known. This paper examines the risk (mean squared error) of the ridge estimator under the constraint  $\theta'\theta \leq c$  and determines optimal values of  $K$  for which the risk is smaller than the risk of the least squares estimators where  $c$  is a constant.

Key words and phrases: Regression, Ridge Analysis, Multicollinearity, Mean Squared Error

AMS Classification: 62J05

ACCESSION for		
NTIS	White Section	<input checked="" type="checkbox"/>
DDC	Buff Section	<input type="checkbox"/>
UNANNOUNCED		<input type="checkbox"/>
JUSTIFICATION _____		
BY _____		
DISTRIBUTION/AVAILABILITY CODES		
Dist.	AVAIL.	and/or SPECIAL
A		

\* The author's work was supported by The Office of Naval Research under Contract N00014-75-C-0451

## RIDGE ESTIMATION IN LINEAR REGRESSION

James S. Hawkes  
Department of Mathematical Sciences  
Clemson University

1. Introduction. In applications of multiple linear regression, the explanatory variables under consideration are often interrelated. The relation is technically called multicollinearity or near multicollinearity. The ordinary least squares estimate of the regression coefficients tends to become "unstable" in the presence of multicollinearity. More precisely, the variance of some of the regression coefficients becomes large. Hoerl (1959), (1962) and Hoerl and Kennard (1970,a), (1970,b) have suggested a class of estimators known as ridge estimators as an alternative to the least squares estimation in the presence of multicollinearity. The new method of estimation is called ridge analysis.

Ridge analysis has drawn considerable interest in recent years. The technique has been developed and new results have been obtained by several authors, e.g., Hawkins (1975), Hemmerle (1975), Sidik (1975). Newhouse and Oman (1975) have conducted a series of Monte Carlo experiments to compare the performance of ridge analysis with the least squares.

The ridge analysis is an ad hoc procedure which gives a biased estimator. We compare the ridge estimator with the least squares estimator with respect to the mean squared error. Hoerl and Kennard (1970,a) have claimed in their paper that for a certain choice of a parameter ( $K$ ) the ridge estimator is uniformly superior to the least squares estimator. This is not true. It appears that the limitation of any optimal property of the ridge estimator and its relation to other known estimators is not often clearly comprehended by many applied statisticians engaged in regression analysis. The object of this paper is to expose the essential features of ridge analysis.

In the following section we show a basis for the choice of the ridge estimator, its biased character and compare it with an unbiased estimator. Furthermore, we obtain conditions under which the ridge estimator has smaller mean squared error than the least squared estimator.

2. Ridge analysis. Consider the linear regression model

$$Y = X\theta + \epsilon$$

where  $Y$  is  $n \times 1$  vector of observations,  $X$  is  $n \times p$  design matrix,  $\theta$  is  $p \times 1$  vector of unknown parameters  $\epsilon$  is  $n \times 1$  vector of the observational errors. It is assumed that the components of  $\epsilon$  are uncorrelated, and have a common variance equal to  $\sigma^2$ , say. Also,  $E(\epsilon) = 0$ . Let prime denote the transpose of a vector or matrix. The least squares estimate of  $\theta$  is obtained by minimizing  $(Y - X\theta)'(Y - X\theta)$  with respect to  $\theta$ , and is given by

$$(2.1) \quad \hat{\theta} = (X'X)^{-1} X'Y.$$

It is assumed that the columns of  $X$  are linearly independent and therefore the rank of the design matrix  $X'X$  is equal to  $p$ .

We have  $E[\hat{\theta}] = \theta$ . That is,  $\hat{\theta}$  is an unbiased estimator of  $\theta$ . Let  $\lambda_1, \dots, \lambda_p$  denote the characteristic roots of  $X'X$ . The mean squared error of  $\hat{\theta}$  ( $MSE\hat{\theta}$ ) is given by

$$(2.2) \quad \begin{aligned} E(\hat{\theta} - \theta)'(\hat{\theta} - \theta) &= \sigma^2 \text{trace } (X'X)^{-1} \\ &= \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i}. \end{aligned}$$

If the explanatory variables are nearly multicollinear then the matrix  $X'X$  is illconditioned, that is, one (or more) of the characteristic roots of  $X'X$  is small. In that case  $MSE\hat{\theta}$  becomes large, as it is seen from (2.2). We can avoid  $MSE\hat{\theta}$  becoming large by inflating the characteristic roots. That is, substituting  $\tilde{\theta}$  for  $\hat{\theta}$ , given by

$$(2.3) \quad \tilde{\theta} = (X'X + KI)^{-1} X'Y$$

where  $I$  is  $p \times p$  identity matrix and  $K$  is a positive number. The estimator  $\tilde{\theta}$  is the ridge estimator, proposed by Hoerl and Kennard as an alternative to the least squares estimator.

We have

$$(2.4) \quad E\tilde{\theta} = (X'X + KI)^{-1} (X'X)\theta$$

Therefore,  $\tilde{\theta}$  is a biased estimator of  $\theta$  unless  $K = 0$ , in which case  $\tilde{\theta} = \hat{\theta}$ . Let  $P$  be an orthogonal matrix diagonalizing  $X'X$ , that is

$$PX'XP' = D$$

where  $D$  is a diagonal matrix with the  $i$ th diagonal element equal to  $\lambda_i$ . Let  $\alpha = (\alpha_1, \dots, \alpha_p)' = P\theta$ . The mean squared error is given after simplification by

$$(2.5) \quad E(\tilde{\theta} - \theta)'(\tilde{\theta} - \theta) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + K)^2} + K^2 \sum_{i=1}^p \frac{\alpha_i^2}{(\lambda_i + K)^2}.$$

From (2.2) and (2.5) we see that for any given  $K > 0$

$$MSE\tilde{\theta} > MSE\hat{\theta}$$

for sufficiently large values of  $\theta'\theta$ . Therefore, the ridge estimator can be compared to the least squares estimator only if  $\theta$  is constrained. Suppose it is known apriori that  $\theta'\theta \leq c$  where  $c$  is a positive number. This condition would be

realized in many practical situations. Since  $\theta' \theta = \alpha' \alpha$ , we have  $\alpha_i^2 \leq c$ ,  $i = 1, \dots, p$ . Hence from (2.5) we get

$$(2.6) \quad \text{MSE } \tilde{\theta} \leq \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + K)^2} + K^2 c \sum_{i=1}^p \frac{1}{(\lambda_i + K)^2}.$$

Theorems 2.1, 2.2, and 2.3 below give certain results on the choice of  $K$  in order that the ridge estimator has smaller mean squared error than the least squares estimator.

Theorem 2.1. If  $\theta' \theta \leq c$  then  $\text{MSE } \tilde{\theta} < \text{MSE } \hat{\theta}$  for  $0 < K \leq \frac{2\sigma^2}{c}$ .

Proof: From (2.6) we have for  $K \leq \frac{2\sigma^2}{c}$

$$\begin{aligned} \text{MSE } \tilde{\theta} &\leq \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + K)^2} + \frac{2K}{(\lambda_i + K)^2} \\ &= \sigma^2 \sum_{i=1}^p \frac{\lambda_i + 2K}{(\lambda_i + K)^2} \\ &< \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i} \\ &= \text{MSE } \hat{\theta}. \end{aligned}$$

Theorem 2.2. If  $\theta' \theta < \frac{\sigma^2}{p} \sum_{i=1}^p \frac{1}{\lambda_i}$  then  $\text{MSE } \tilde{\theta} < \text{MSE } \hat{\theta}$  for  $K > 0$ .

Proof: Let  $D(K)$  denote the quantity on the right hand side of (2.6). Differentiating  $D(K)$  with respect to  $K$  we have

$$(2.7) \quad \partial D(K) / \partial K = \sum_{i=1}^p \frac{2\lambda_i (cK - \sigma^2)}{(\lambda_i + K)^3}.$$

The right hand side of (2.7) is equal to zero for  $K = \frac{\sigma^2}{c}$  and is  $<(>) 0$  for  $K <(>) \frac{\sigma^2}{c}$ . Therefore,  $D(K)$  is first decreasing then increasing as  $K$  varies from 0 to  $\infty$ . Now

$$\begin{aligned} D(0) &= \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i} \\ &= \text{MSE } \hat{\theta} \end{aligned}$$

$$D(\infty) = pc.$$

Hence

$$D(c) \leq \max(p\sigma^2, \text{MSE } \hat{\theta})$$

$$= \text{MSE } \hat{\theta} \text{ for } c \leq \frac{\sigma^2}{p} \sum_{i=1}^p \frac{1}{\lambda_i}.$$

Since  $D(K)$  is an upper bound on the value of  $\text{MSE } \hat{\theta}$ , the theorem follows.  $\square$

From a Bayesian point of view suppose that the components of  $\theta$  are independently and identically distributed with means  $\xi$  and variance  $\tau^2$ .

Theorem 2.3. If the components of  $\theta$  are independently and identically distributed with mean  $\xi$  and variance  $\tau^2$  then the average mean squared error of the ridge estimator is minimized for  $k = \sigma^2 / (\xi^2 + \tau^2)$ .

Proof: Let  $E$  denote expectation with respect to the given prior distribution of  $\theta$ . We have

$$(2.8) \quad E(\text{MSE } \hat{\theta}) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + K)^2} + K^2 \sum_{i=1}^p \frac{\xi^2 + \tau^2}{(\lambda_i + K^2)}.$$

As in the proof of Theorem 2.2 we find that the right hand side of (2.8) is minimized for  $K = \sigma^2 / (\xi^2 + \tau^2)$ .  $\square$

We have seen above that the ridge estimator is preferred to the least squares estimator in certain situations when the parameter  $\theta$  is constrained. The following theorem gives a basis for the choice of the ridge estimator under the given constraint.

Theorem 2.4. The value of  $\theta$  minimizing  $R(\theta) = (Y - X\theta)'(Y - X\theta)$ , given  $\theta'\theta \leq c$  is equal to  $\hat{\theta}$  where  $K$  is chosen such that  $\hat{\theta}'\hat{\theta} = c$ .

Proof: By direct computation we get

$$(2.9) \quad \tilde{\theta}'\tilde{\theta} = (PX'Y)'(D + KI)^{-2}(PX'Y).$$

It is seen from (2.9) that  $\tilde{\theta}'\tilde{\theta}$  is decreasing in  $K$ . Therefore, the value of  $K$ , given by  $\tilde{\theta}'\tilde{\theta} = c$  is uniquely determined.

We have

$$(2.10) \quad R(\hat{\theta}) = (Y - X\hat{\theta})'(Y - X\hat{\theta})$$

$$= (Y - X\hat{\theta})'(Y - X\hat{\theta}) + (X'Y)'[(X'X + KI)^{-1} - (X'X)^{-1}]$$

$$X'X[(X'X + KI)^{-1} - (X'X)^{-1}](X'Y)$$

$$= (Y - X\hat{\theta})'(Y - X\hat{\theta}) + (PX'Y)' D^* (PX'Y)$$



where  $D^*$  is a  $p \times p$  diagonal matrix whose  $i$ th diagonal element is equal to

$$\frac{k^2}{\lambda_i^2 (k + \lambda_i)^2}$$

From (2.10) we see that  $R(\theta)$  is increasing in  $k$ .

Consider the problem of minimizing  $R(\theta)$  with respect to  $\theta$  under the constraint  $\theta'\theta = c$ . By the Lagrangian method the minimizing value of  $\theta$  is given by

$$\lambda \theta - X'(Y - X\theta) = 0$$

or

$$\theta = (X'X + \lambda I)^{-1} X'Y$$

where  $\lambda$  is determined such that  $\theta'\theta = c$ . That is, the minimizing value of  $\theta$  is the ridge estimator  $\tilde{\theta}$ , where  $k$  is determined such that  $\tilde{\theta}'\tilde{\theta} = c$ .

We have shown above  $\tilde{\theta}'\tilde{\theta}$  is decreasing in  $k$  and that  $R(\tilde{\theta})$  is increasing in  $k$ . It follows that  $\tilde{\theta}$  which  $k$  minimizes  $R(\theta)$  under the constraint  $\theta'\theta = c$ , minimizes  $R(\theta)$  also under the constraint  $\theta'\theta \leq c$ .  $\square$

Remark: We have a comparison between the least squares estimation and ridge estimation. The ridge estimator is given by minimizing  $R(\theta)$  under a certain constraint on the value of  $\theta'\theta$ , whereas the least squares estimator is given by minimizing  $R(\theta)$  without that constraint.

Throughout the foregoing discussion we have assumed that the quantity  $k$  arising in the definition of the ridge estimator  $\tilde{\theta}$ , is a scalar constant. By letting  $k$  depend on the observation  $Y$  suitably, we might be able to obtain an estimator which has a smaller MSE than the least squares estimator for all values of  $\theta$ . Hoerl and Kennard (1970, a) have suggested an iterative method of choosing such a value of  $K$ . However, they did not show that the final estimator had a smaller MSE than the least squares estimator. On the other hand, (see Alam (1974)) any estimator of the form

$$\phi((Y'X(X'X)^{-1}X'Y)/\sigma^2) \tilde{\theta}$$

has smaller MSE than  $\tilde{\theta}$  where  $\phi(Z)$  is a function, such that,  $Z(1-\phi(Z))$  is non-decreasing in  $Z$  and  $0 \leq Z(1-\phi(Z)) \leq 2\alpha p - 4$  and

$$\alpha = \left( \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i} \right) \min(\lambda_1, \dots, \lambda_p).$$

See also, Sclove (1968) and Stein (1960).

## References

- [1] Alam, K. (1974). Minimax estimators of the mean of a multivariate normal distribution. Tech. Report, Mathematical Sciences, Clemson University.
- [2] Hawkins, D. M. (1975). Relations Between Ridge Regression and Eigenanalysis of the Augmented Correlation Matrix. *Technometrics*, Vol. 17, pp. 477-80.
- [3] Hemmerle, W. J. (1975). An Explicit Solution for Generalized Ridge Regression. *Technometrics*, Vol. 17, pp. 309-14.
- [4] Hoerl, A. E. (1959). Optimum Solution of Many Variable Equations. *Chemical Engineering Progress* 55, pp. 69-78.
- [5] Hoerl, A. E. (1962). Applications of Ridge Analysis to regression problems. *Chemical Engineering Progress* 58, pp. 54-59.
- [6] Hoerl, A. E. and Kennard, R. W. (1970). Biased Estimation for Nonorthogonal Problems. *Technometrics*, Vol. 12, pp. 55-67.
- [7] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Application to non-orthogonal problems. *Technometrics*, Vol. 12, pp. 69-82.
- [8] Newhouse, J. P. and Oman, S. D. (1971). An evaluation of ridge estimators. Rand Technical Report R-716-PR.
- [9] Sclove, S. L. (1968). Improved estimators for coefficients in linear regression. *J. Amer. Statisti. Assoc.* 63, 597-606.
- [10] Sidik, S. M. (1975). Comparison of Some Biased Estimation Methods (Including Ordinary Subset Regression) in the Linear Model. Nasa Technical Report TN D-7932.
- [11] Stein, C. M. (1960). Multiple regression. Contributions to probability and statistics. Essays in honour of Harold Hotelling, Olkin, I (ed.), Stanford University Press, 423-43.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER N 80	2. GOVT ACCESSION NO. AD-A083 521	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Ridge Estimation In Linear Regression		5. TYPE OF REPORT & PERIOD COVERED TR #232
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) James S. Hawkes		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS Clemson University Dept. of Mathematical Sciences Clemson, South Carolina 29631		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 047-202 NR 042-271
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Code 436-434 Arlington, Va. 22217		12. REPORT DATE 10/76
		13. NUMBER OF PAGES 8
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/ DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report)  Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)  Regression, Ridge Analysis, Multicollinearity, Mean Squared Error		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Consider the linear regression model $Y = X\theta + \epsilon$ . Recently, a class of estimators, variously known as ridge estimators, has been proposed as an alternative to the least squares estimators in the case of collinearity, that is, when the design matrix $X'X$ is nearly singular. The ridge estimator is given by $\hat{\theta} = (X'X + KI)^{-1} X'Y$ , where $K$ is a constant to be determined. An optimal choice of the value of $K$ is not known. This paper examines the risk (mean squared error) of the ridge estimator under the constraint $\theta'\theta \leq c$ and determines optimal values of $K$ for which		

DD FORM 1473

EDITION OF 1 NOV 55 IS OBSOLETE  
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

the risk is smaller than the risk of the least squares estimators where  $c$  is a constant.